USC EE599 RL and LLMs 2025 Fall - Final Project Report Judging Model for Large-scale Multimodality Benchmarks

Min-Han Shih, Yu-Hsin Wu, Yu-Wei Chen
Department of Electrical and Computer Engineering,
Viterbi School of Engineering, University of Southern California, United States
{minhansh, yuhsinwu, ychen543}@usc.edu

Abstract

We propose a dedicated multimodal Judge Model designed to provide reliable, explainable evaluation across a diverse suite of tasks. Our benchmark spans text, audio, image, and video modalities, drawing from carefully sampled public datasets with fixed seeds to ensure reproducibility and minimize train–test leakage. Instead of simple scoring, our framework aggregates multimodal judgments, analyzes the quality and reasoning consistency of model outputs, and generates diagnostic feedback. We evaluate state-of-the-art MLLMs, including Gemini-2.5-flash, Phi-4, and Qwen-2.5-omni, across 280 multimodal samples and compare judge-model assessments with human annotators. Results show strong alignment between the Judge Model and human scores, demonstrating its potential as a scalable, interpretable evaluation pipeline for future multimodal AI research.

1 Introduction

Recent advances in Large Language Models (LLMs) and Multimodal Large Language Models (MLLMs) have demonstrated remarkable progress in understanding and reasoning across multiple modalities such as text, image, audio, and video. Models like GPT-5 (OpenAI-Team, 2025), Gemini 2.5 (Gemini-Team, 2025b,a), and Claude 4 (Anthropic-Team, 2025) sonnet can now interpret complex multimodal inputs, perform reasoning over long contexts, and generate structured, contextually aware outputs. However, as the capability and diversity of MLLMs expand, objectively and efficiently evaluating their performance has become an increasingly challenging problem.

Traditional evaluation pipelines rely on either human annotation or static benchmark scoring, which are often time-consuming, inconsistent, and limited to task-specific performance. Meanwhile, recent approaches, such as LLM-as-a-Judge, leverage large models as evaluators. (Calderon et al., 2025; Dorner et al., 2025; D'Souza et al., 2025) Although these methods reduce human workload, they introduce several issues: (1) lack of multimodal understanding, since many judge systems are purely text-based; and (2) insufficient explainability and reliability when compared to human judgment on complex reasoning or cross-modal alignment tasks.

To address these limitations, our project proposes a **Judge Model** designed specifically for large-scale multimodal benchmarks. Instead of scoring, our framework aggregates and calibrates judgments across multiple perspectives. The Judge will evaluate the outputs of various MLLMs, such as Gemini 2.5 (Gemini-Team, 2025a), Phi-4 (Abdin et al., 2024), and Qwen-2.5 (Qwen et al., 2025), analyzing the quality of the response, the alignment of the modality and the consistency of the reasoning trace. Beyond scoring, it will also generate comprehensive feedback and guidance, aiming to establish an interpretable and reliable evaluation pipeline for multimodal AI systems.

2 Related Work

2.1 Reasoning LLMs

LLMs know when to think step-by-step and only generate step-by-step reasoning when necessary. (Wei et al., 2023) Prior works on Chain-of-Thought (CoT) mostly focus on the correctness of the CoT reasoning steps or whether the reasoning steps are faithful to the question and support the final answer. (Ye and Durrett, 2022; Golovneva et al., 2023)

2.2 LLM-as-the-Judge

LLM-based evaluators are LLMs that are prompted to judge the quality of some samples based on specific criteria. LLM-based evaluators evaluate the quality of a single sample using a score such as Likert scores (Likert, 1932) on a scale of 1 to 5. (Chiang et al., 2024; Chiang and Lee, 2023b,a, 2024; Zheng et al., 2023) LLM-based evaluators can also compare the quality of a pair of samples and judge which one is better. (Zheng et al., 2023)

2.3 Multimodal Benchmark

Vision benchmark expands from image benchmarks to video benchmarks that introduce temporal structure and richer scene context (Nguyen et al., 2025), such as Vision Quesion-Answering (Agrawal et al., 2016; Goyal et al., 2017), chart understanding (Masry et al., 2022; Wang et al., 2024b; Mathew et al., 2021), and general capability suites (Fu et al., 2025; Li et al., 2023; Liu et al., 2024a). Different video datasets emphasize different abilities: temporal ordering (Cai et al., 2024; Liu et al., 2024b), procedural understanding (Tang et al., 2019; Xiao et al., 2021), egocentric perception (Grauman et al., 2022; Damen et al., 2018), and world modeling (Hong et al., 2025). Audio benchmarks such as AudioSet (Kim et al., 2019) and VGGSound (Chen et al., 2020) target event-level acoustic classification, while speech datasets cover specific facets of communication: VoxCeleb (Nagrani et al., 2017; Chung et al., 2018) for speaker identity, AudioCaps / SpeechCaps (Kim et al., 2019; Huang et al., 2025b) for audio/speech captioning, LRS (Afouras et al., 2018) for transcription, and AVA-ActiveSpeaker (Kim et al., 2021) for frame-level speaking and audibility labels. Dynamic-SUPERB (Huang et al., 2024, 2025a) proposed a dynamic, collaborative benchmark in speech and audio tasks.

3 Datasets

3.1 Text-to-Text Datasets

To build a compact, reproducible benchmark, we pull small, fixed-size samples from widely used public datasets on HuggingFace Hub. For each dataset we explicitly name the split used so that readers can assess the risk of train–test leakage in future model training. We sample without replacement using a fixed random seed. The data source and the number of each task are listed in Table 1.

Task Family	Dataset (HF ID)	Split	#Items	Answer Type
Reasoning–Code	`nvidia/OpenCodeReasoning`	`split_0`	20	code / final string
Reasoning–Math	`nvidia/OpenMathReasoning`	`cot`	20	generative (final value)
Expert MCQ	`Idavidrein/gpqa` (*fallback:* `fingertap/GPQA-Diamond`)	`test`	10	multiple choice
Reading Comprehension	`ucinlp/drop`	`validation`	20	extractive span
Commonsense Reasoning	`Rowan/hellaswag`	`validation`	15	multiple choice
Commonsense Reasoning	`jet-ai/social_i_qa`	`validation`	15	multiple choice
Instruction Following	`google/IFEval`	`train`	15	generative (constrainted)
Instruction Following	`YuxinJiang/FollowBench`	`train` (*fallback:* `validation`)	15	generative (constrainted)
Total			**130**	

Table 1: Datasets, splits, counts, and answer types used in our text benchmark.

3.2 Text-to-Text Tasks

1 .**Reasoning–Code:** program synthesis problem solving requiring structured outputs or final values. (Ahmad et al., 2025)

2. **Reasoning–Math:** multi-step mathematical reasoning aiming at a final canonical answer. (Moshkov et al., 2025)

3. **Expert MCQ:** graduate-level STEM multiple-choice questions (Rein et al., 2023).

4. **Reading Comprehension:** passage-based QA with extractive answers (Dua et al., 2019)

5. **Commonsense Reasoning:** everyday physical/-social reasoning in multiple-choice format. (Zellers et al., 2019; Sap et al., 2019)

6. **Instruction Following:** adherence to explicit constraints and formatting instructions. (Jiang et al., 2023; Zhou et al., 2023)

3.3 Audio-Language Datasets

Our audio-language datasets are constructed from four open-source datasets hosted on Hugging Face. These datasets were carefully selected to ensure diversity across audio modalities and task types, including environmental sound understanding, bird sound classification, music generation, and speech question answering. The sources and task types are summarized in Table 2.

Task Family	Dataset (HF ID)	Split	#Items
Audio Captioning	`kuanhuggingface/audiocaps_hallucination`	`test`	10
Birdsound Detection	`tglcourse/5s_birdcall_samples_top20`	`validation`	10
Music Style Detection	`nyuuzyou/suno`	`train`	10
Speech Question-Answering	`AudioLLMs/public_sg_speech_qa_test`	`test`	20
Total			**50**

Table 2: Datasets, splits, counts, and answer types used in our audio benchmark.

3.4 Audio-Language Tasks

1. **Question–Answer:** Involve recognizing and describing the presence or absence of specific sounds, or answering questions based on audio content. (Kim et al., 2019; Wang et al., 2025a)

2. **Instruction–Following:** Evaluate the model's ability to interpret and execute textual instructions for auditory inputs. (Wang et al., 2025a)

3.5 Video Datasets

Our video dataset comes from the Hugging Face public dataset. It consists of AI-generated short clips created by generative video models such as Sora, Veo2, and Kling. Each clip is paired with human-curated questions designed to test the reasoning of a model about common sense and physical consistency. To ensure reproducibility, we sample using a fixed random seed and select a total of 50 video clips from non-training sets, which minimizes the chance of overlap with materials that LLMs might have already seen during pretraining. This subset provides a controlled and diverse

evaluation set for testing video content understanding and hallucination robustness in MLLMs. The sources and task types are summarized in Table 3.

Dataset (HF ID)	Split	#Items
IntelligenceLab/VideoHallu	validation	25
IntelligenceLab/VideoHallu	test	25
Total		**50**

Table 3: Datasets, splits, counts, and answer types used in our video benchmark.

3.6 Video Tasks

Our task focuses on common sense and physical reasoning rather than traditional visual recognition, due to the use of AI-generated synthetic videos. The model is required to analyze each short clip, interpret the given question, and determine whether the depicted scene is plausible or abnormal. By framing the problem in this way, we shift the emphasis from simple visual perception to critical reasoning, allowing us to evaluate how effectively MLLMs can align visual information with their internal understanding of real-world knowledge.

3.7 Iamge Datasets

To construct a diverse and interpretable benchmark for image–language understanding, we curated samples from four public datasets on Hugging Face, each representing a distinct subdomain of visual reasoning. For all datasets, we sample small, fixed-size subsets using a consistent random seed for reproducibility. Only evaluation or test splits are used to minimize potential data leakage from pretraining corpora. The sources and task types are summarized in Table 4.

Task	Dataset (HF ID)	Split	#Items
Image Captioning	nyu-visionx/CV-Bench	test	10
Image Captioning	nirajandhakal/realworldqa	test	10
Chart Understanding	ChartFoundation/ECDBench	test	15
Math Reasoning	MathLLMs/MathVision	test	15
Total			**50**

Table 4: Datasets, splits, counts, and answer types used in our image benchmark.

3.8 Iamge Tasks

1. **Image Captioning:** Evaluates the model's ability to perform commonsense reasoning and holistic scene understanding. It assesses whether the model can generate captions that reflect physical plausibility, spatial and contextual consistency, and realistic visual interpretation. The images span both everyday real-world scenes and complex scenarios that require object recognition, contextual inference, and higher-level reasoning. (Zhu et al., 2025)
2. **Chart Understanding:** Measures quantitative and relational reasoning from structured visuals like bar charts, requiring comparison of trends and numerical relationships. (Yang et al., 2025)
3. **Math Reasoning:** Assesses symbolic and spatial reasoning using mathematical diagrams and geometric figures that demand parsing and quantitative inference. (Wang et al., 2024a, 2025b)

4 Methodology

4.1 Overview of Evaluation Pipeline

We design a scalable automated evaluation pipeline that assesses the capabilities of MLLMs across text, audio, image and video tasks. As illustration in Fig. 1, there are three main components: Multimodal data, Tested MLLMs, and Judge model. For each instance, the system retrieves a sample containing a query and its corresponding context. These inputs are fed into the tested MLLM, which generates a text-only response and a justification.

While our framework is architecturally designed to support a feedback loop–where errors are returned to the model for regeneration–this project focuses on validating the reliability of the Judge Model. Therefore, in our experiments, the pipeline works as a comprehensive forward-pass evaluation. That is, the Judge Model assesses the generated outputs against ground truth references and the multimodal evidence, and gives its own feedback.

4.2 Tested MLLMs

To ensure our benchmark covers a diverse range of capabilities, we evaluate three MLLMs that represent different model families and sizes.

- **Gemini-2.5**: An efficient multimodal model optimized for speed and reasoning.
- **Phi-4**: A powerful open-weight model developed by Microsoft, capable of handling complex tasks.
- **Qwen-2.5**: An omni-modal model designed to process and reason over various modality inputs.

4.3 Judge Model

The core of our methodology is the Judge model, powered by **Gemini-3-pro**. Unlike traditional text-only judges, this agent is explicitly prompted to perform multimodal reasoning. It receives the original task instruction, the multimodal context, the

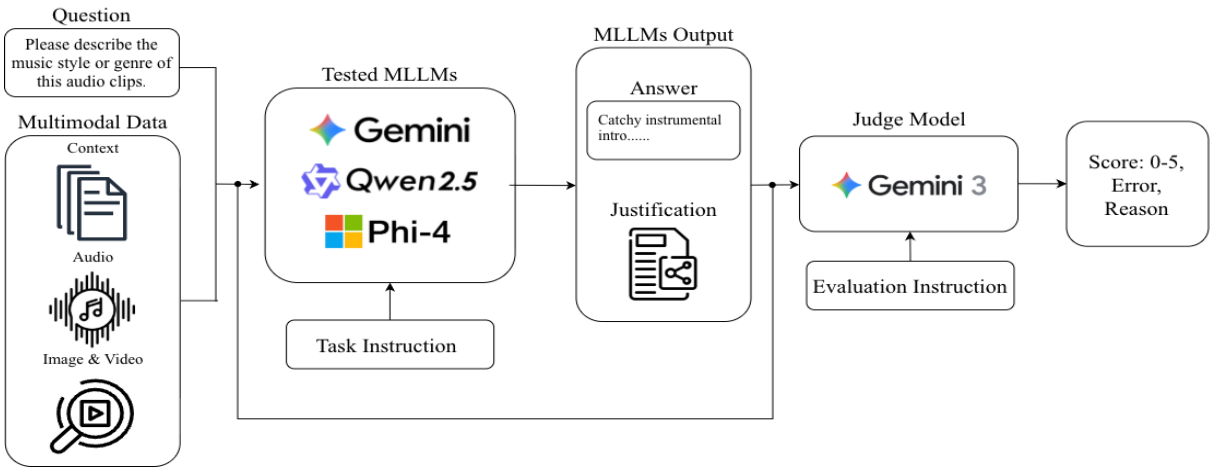

Figure 1: System architecture of the MLLMs evaluation framework. The framework takes multimodal inputs and task instructions to query multiple tested MLLMs. Each model produces an answer along with its justification. A judge model, following evaluation instructions, assesses the responses and outputs scores (0–5) and error analyses.

ground truth, and the tested model's output. The Judge model has three primary objectives:

1. **Response Quality Analysis:** It determines if the final answer matches the ground truth or a physically observable reality.
2. **Reasoning Verification:** It analyzes the "Justification" provided by the tested model to check for logical consistency and hallucination.
3. **Diagnostic Feedback:** Instead of a simple binary pass/fail, the agent outputs a structured evaluation containing a numerical score (0-5), the specific error type, and a detailed reason for the score.

4.4 Evaluation Metrics

To quantify model performance, we utilize a rigorous Likert-scale scoring rubric. The Judge model assigns a score from 0 to 5 based on the following criteria defined in Table 5. Beyond numerical scoring, the protocol includes a qualitative error analysis. As shown in our case studies, the Judge identifies specific failure modes, such as visual hallucinations or auditory omissions.

Score	Description
5	Correct answer with sound justification.
4	Mostly correct answer and justification and minor errors.
3	Partially correct answer and justification, but errors exist.
2	Partially correct answer but unreasonable justification.
1	Mostly wrong answer with unreasonable justification.
0	Totally wrong answer or empty response.

Table 5: Evaluation metrics for the judge model and human judge to quantify tested MLLMs performance.

4.5 Experiment Results

Table 6 summarizes the average scores assigned by three human annotators and our Judge Model across four modalities and three MLLMs. Overall, we observe that tasks involving audio and video tend to achieve higher scores than those involving text-only or image inputs when averaging across all models, suggesting that current MLLMs may already possess relatively mature capabilities for temporal and acoustic understanding under our benchmark setting. In contrast, image-based and especially text-only tasks expose larger performance gaps between models: while the strongest system maintains scores close to 4, weaker models often fall near or below 2, indicating substantial room for improvement in fine-grained reading comprehension and visually grounded reasoning.

From a model-wise perspective, Gemini-2.5 exhibits the most robust cross-modal performance. Its scores remain consistently high across all four modalities in most cases, with human and Judge ratings clustered roughly between 3.8 and 4.5, and without any catastrophic failure mode in a particular modality. This pattern suggests that Gemini's internal representations transfer well between text, audio, image, and video inputs, and that its justifications are generally judged correct and well-grounded. In other words, under our evaluation protocol, Gemini behaves like a genuinely all-round multimodal model rather than a specialist tuned on a single dominant modality.

In contrast, Phi-4 and Qwen-2.5 display more pronounced modality-specific weaknesses. Phi-4 delivers mid-range performance on audio and video, but its scores drop sharply on image tasks, where hallucinated visual details and misinterpreted chart elements frequently lead to scores near 2. Qwen-2.5 shows a complementary pattern: it performs

competitively in audio and video, but its text-only scores are the lowest among all models, reflecting frequent failures in straightforward extraction and reasoning, including false refusals on questions with explicitly stated answers. Overall, these results indicate that cross-modal robustness is far from uniform across current MLLMs: some models generalize smoothly across modalities, while others remain highly sensitive to the dominant modality and the type of reasoning required.

		Human 1	Human 2	Human 3	Judge
Text	*Gemini-2.5*	3.98	**4.44**	4.12	4.28
	Phi-4	2.24	2.12	2.98	2.45
	Qwen-2.5	1.37	1.28	1.5	1.15
Audio	*Gemini-2.5*	3.58	3.62	**4.16**	4.04
	Phi-4	2.66	3.26	3.30	2.82
	Qwen-2.5	3.26	3.78	3.56	3.38
Video	*Gemini-2.5*	3.70	4.10	4.20	**4.54**
	Phi-4	2.78	3.60	3.50	3.12
	Qwen-2.5	2.82	3.50	3.02	3.06
Image	*Gemini-2.5*	3.46	3.82	**3.98**	3.94
	Phi-4	2.20	2.66	1.80	2.24
	Qwen-2.5	2.52	3.28	2.28	2.32

Table 6: Evaluation averaged score of Gemini-2.5, Phi-4, and Qwen-2.5 by three human annotators and the judge model, Gemini-3-pro, across multi-modal tasks.

5 Analysis

5.1 Score Distribution

Figure 2 visualizes the score distributions of the three human annotators and the Judge Model for each model, complementing the averaged scores reported in Table 6. Across Gemini-2.5, Phi-4, and Qwen-2.5, we observe that the Judge's curve closely tracks the human trends in all four modalities: systems that receive higher mean scores from humans also receive higher scores from the Judge, and the relative ordering of models is preserved within each modality. At the same time, the Judge is slightly more conservative in its absolute calibration—its scores are typically lower than the three-annotator average by about 0.1–0.3 points. This systematic offset reflects the Judge's stricter penalty on inconsistencies between the final answer and the provided justification, such as partially correct answers supported by hallucinated or logically incomplete reasoning. Despite this mild negative bias, the near-perfect agreement in rank ordering suggests that our automatic evaluator is reliable for comparative assessment of MLLMs, even if small corrections may be needed when interpreting its Likert scores in isolation.

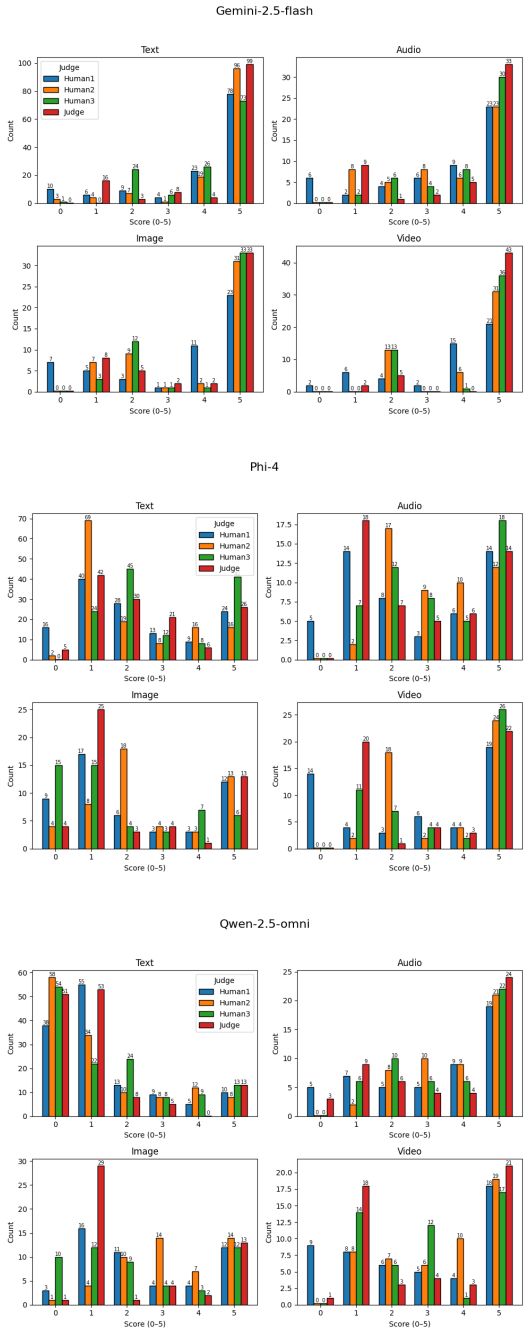

Figure 2: Human vs. Judge alignment across MLLMs.

5.2 Case Study

5.2.1 Text

We illustrate a representative text QA example where the model must identify Carolina's quarterback from a short American-football game recap. The paragraph explicitly states that "QB Jake Delhomme completed a 3-yard TD pass to WR Muhsin Muhammad," and the question is "who is carolina's quarterback?". As shown in Table 7, Gemini-2.5 correctly extracts *Jake Delhomme* and receives a perfect score from the Judge. In contrast, Qwen-2.5

incorrectly claims that the question cannot be answered, even though the answer is explicitly stated in the context, leading the Judge to label this case as a *false refusal* with a low score. This example highlights that our Judge not only checks final accuracy but also penalizes missed evidence when the ground-truth answer is observable in the input.

Model	Prediction	Score	Error Type
Gemini-2.5	Jake Delhomme	5	None
Qwen-2.5	No, it's impossible to answer this question.	1	False refusal

Table 7: Text QA case study: identifying Carolina's quarterback from a game recap paragraph.

5.2.2 Image

For the image case study, we consider a chart-understanding question where the model must read the x-axis of a synthetic timeline plot shown in Fig. 3 and answer: "How many labeled epochs are present on the x-axis, and what are their names?". The ground-truth chart contains four epochs: *Ancient Algorithms*, *Medieval Methods*, *Renaissance Techniques*, and *Modern Simulations*. As summarized in Table 8, Gemini-2.5 correctly recovers both the count and all four labels, and is therefore assigned the maximum score by the Judge. In contrast, Phi-4 also predicts that there are four epochs but hallucinates an entirely different set of names (e.g., *Early Adopters*, *Historical Adopters*), which do not appear anywhere in the chart. The Judge accordingly treats this as a hallucination case and assigns only partial credit. This example shows that our evaluator is sensitive not only to coarse numerical correctness but also to fine-grained semantic grounding in the visual evidence.

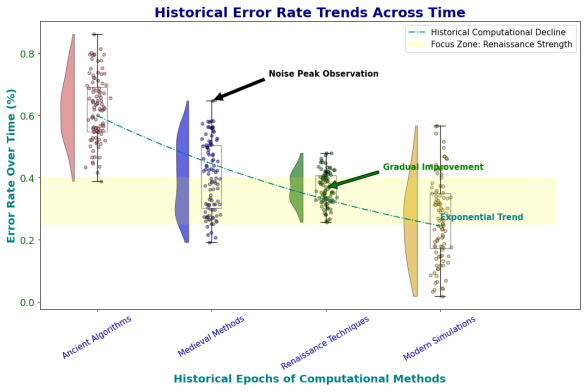

Figure 3: The input chart example for image case study.

6 Future Work

A promising direction for future work is to repurpose our Judge outputs as supervision signals. Each instance already includes a scalar score, an error type, and a natural-language explanation, which together form a rich signal that can be used as a reward model for reinforcement learning from LLM feedback or as a set of soft constraints in a Constitutional AI framework. Rather than stopping at offline evaluation, we could close the loop by fine-tuning MLLMs to maximize Judge scores and reduce specific failure modes such as hallucination and false refusal. Extending this idea beyond single-turn QA, future work may apply the same feedback mechanism to multi-turn agentic settings, where the Judge evaluates entire interaction traces instead of only final answers. Such a scaled-up pipeline would turn our Judge from static into active, enabling dynamically "learning from LLMs".

Model	Prediction	Score	Error Type
Gemini-2.5	There are 4 labeled epochs: Ancient Algorithms, Medieval Methods, Renaissance Techniques, and Modern Simulations.	5	None
Phi-4	The x-axis has 4 labeled epochs: Early Adopters, Historical Adopters, General Adoption, and Late Adopters.	3	Hallucination

Table 8: Image case study: reading the labeled epochs on the x-axis of a chart.

7 Conclusion

We introduced a multimodal benchmark[1] with a Judge for evaluating MLLMs across text, audio, image, and video tasks, combining Likert-scale scores with structured error types and natural-language explanations. Our experiments on Gemini-2.5, Phi-4, and Qwen-2.5, show that the Judge's ratings closely track human annotators while being slightly more conservative, and reliably preserve the relative ranking of systems across modalities. Case studies further demonstrate that the Judge can distinguish between correct, hallucinated, and falsely refused answers in a fine-grained manner.

8 Members Contributions

Task	Min-Han Shih	Yu-Hsin Wu	Yu-Wei Chen
Data Collection	Text	Audio	Image/Video
Model Inference	Gemini	Phi-4	Qwen-2.5
Human Evaluation	✓ (33%)	✓ (33%)	✓ (33%)
Experiment Design	✓ (100%)		
Coding	50%	25%	25%
Presentation Slides	✓ (80%)	✓ (10%)	✓ (10%)
Writing (Draft)	✓ (100%)		
Writing (Revision)		✓ (50%)	✓ (50%)
Visualization Plots		✓ (50%)	✓ (50%)

Table 9: Division of labor among three contributors.

[1] https://github.com/oscar-shih/multimodal_judge_benchmark
https://huggingface.co/multi-judge

References

Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J Hewett, Mojan Javaheripi, Piero Kauffmann, et al. 2024. Phi-4 technical report. *arXiv preprint arXiv:2412.08905*.

Triantafyllos Afouras, Joon Son Chung, and Andrew Zisserman. 2018. Lrs3-ted: a large-scale dataset for visual speech recognition.

Aishwarya Agrawal, Jiasen Lu, Stanislaw Antol, Margaret Mitchell, C. Lawrence Zitnick, Dhruv Batra, and Devi Parikh. 2016. Vqa: Visual question answering.

Wasi Uddin Ahmad, Sean Narenthiran, Somshubra Majumdar, Aleksander Ficek, Siddhartha Jain, Jocelyn Huang, Vahid Noroozi, and Boris Ginsburg. 2025. Opencodereasoning: Advancing data distillation for competitive coding.

Anthropic-Team. 2025. Introducing claude 4.

Mu Cai, Reuben Tan, Jianrui Zhang, Bocheng Zou, Kai Zhang, Feng Yao, Fangrui Zhu, Jing Gu, Yiwu Zhong, Yuzhang Shang, Yao Dou, Jaden Park, Jianfeng Gao, Yong Jae Lee, and Jianwei Yang. 2024. Temporalbench: Benchmarking fine-grained temporal understanding for multimodal video models.

Nitay Calderon, Roi Reichart, and Rotem Dror. 2025. The alternative annotator test for LLM-as-a-judge: How to statistically justify replacing human annotators with LLMs. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16051–16081, Vienna, Austria. Association for Computational Linguistics.

Honglie Chen, Weidi Xie, Andrea Vedaldi, and Andrew Zisserman. 2020. Vggsound: A large-scale audio-visual dataset.

Cheng-Han Chiang, Wei-Chih Chen, Chun-Yi Kuan, Chienchou Yang, and Hung-yi Lee. 2024. Large language model as an assignment evaluator: Insights, feedback, and challenges in a 1000+ student course. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 2489–2513, Miami, Florida, USA. Association for Computational Linguistics.

Cheng-Han Chiang and Hung-yi Lee. 2023a. Can large language models be an alternative to human evaluations? In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15607–15631, Toronto, Canada. Association for Computational Linguistics.

Cheng-Han Chiang and Hung-yi Lee. 2023b. A closer look into using large language models for automatic evaluation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 8928–8942, Singapore. Association for Computational Linguistics.

Cheng-Han Chiang and Hung-yi Lee. 2024. Over-reasoning and redundant calculation of large language models. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 161–169, St. Julian's, Malta. Association for Computational Linguistics.

Joon Son Chung, Arsha Nagrani, and Andrew Zisserman. 2018. Voxceleb2: Deep speaker recognition. In *Interspeech 2018*, pages 1086–1090.

Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. 2018. Scaling egocentric vision: The epic-kitchens dataset.

Florian E. Dorner, Vivian Y. Nastl, and Moritz Hardt. 2025. Limits to scalable evaluation at the frontier: Llm as judge won't beat twice the data.

Jennifer D'Souza, Hamed Babaei Giglou, and Quentin Münch. 2025. YESciEval: Robust LLM-as-a-judge for scientific question answering. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13749–13783, Vienna, Austria. Association for Computational Linguistics.

Dheeru Dua, Yizhong Wang, Pradeep Dasigi, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. 2019. DROP: A reading comprehension benchmark requiring discrete reasoning over paragraphs. In *Proc. of NAACL*.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, Yunsheng Wu, Rongrong Ji, Caifeng Shan, and Ran He. 2025. Mme: A comprehensive evaluation benchmark for multimodal large language models.

Gemini-Team. 2025a. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities.

Gemini-Team. 2025b. Gemini: A family of highly capable multimodal models.

Olga Golovneva, Moya Chen, Spencer Poff, Martin Corredor, Luke Zettlemoyer, Maryam Fazel-Zarandi, and Asli Celikyilmaz. 2023. Roscoe: A suite of metrics for scoring step-by-step reasoning.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the v in vqa matter: Elevating the role of image understanding in visual question answering.

Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, Miguel Martin, Tushar Nagarajan, Ilija Radosavovic, Santhosh Kumar Ramakrishnan, Fiona Ryan, Jayant Sharma, Michael Wray, Mengmeng Xu,

Eric Zhongcong Xu, Chen Zhao, Siddhant Bansal, Dhruv Batra, Vincent Cartillier, Sean Crane, Tien Do, Morrie Doulaty, Akshay Erapalli, Christoph Feichtenhofer, Adriano Fragomeni, Qichen Fu, Abrham Gebreselasie, Cristina Gonzalez, James Hillis, Xuhua Huang, Yifei Huang, Wenqi Jia, Weslie Khoo, Jachym Kolar, Satwik Kottur, Anurag Kumar, Federico Landini, Chao Li, Yanghao Li, Zhenqiang Li, Karttikeya Mangalam, Raghava Modhugu, Jonathan Munro, Tullie Murrell, Takumi Nishiyasu, Will Price, Paola Ruiz Puentes, Merey Ramazanova, Leda Sari, Kiran Somasundaram, Audrey Southerland, Yusuke Sugano, Ruijie Tao, Minh Vo, Yuchen Wang, Xindi Wu, Takuma Yagi, Ziwei Zhao, Yunyi Zhu, Pablo Arbelaez, David Crandall, Dima Damen, Giovanni Maria Farinella, Christian Fuegen, Bernard Ghanem, Vamsi Krishna Ithapu, C. V. Jawahar, Hanbyul Joo, Kris Kitani, Haizhou Li, Richard Newcombe, Aude Oliva, Hyun Soo Park, James M. Rehg, Yoichi Sato, Jianbo Shi, Mike Zheng Shou, Antonio Torralba, Lorenzo Torresani, Mingfei Yan, and Jitendra Malik. 2022. Ego4d: Around the world in 3,000 hours of egocentric video.

Jack Hong, Shilin Yan, Jiayin Cai, Xiaolong Jiang, Yao Hu, and Weidi Xie. 2025. Worldsense: Evaluating real-world omnimodal understanding for multimodal llms.

Chien-Yu Huang, Wei-Chih Chen, Shu wen Yang, Andy T. Liu, Chen-An Li, Yu-Xiang Lin, Wei-Cheng Tseng, Anuj Diwan, Yi-Jen Shih, Min-Han Shih, Jiatong Shi, William Chen, Chih-Kai Yang, Wenze Ren, Xuanjun Chen, Chi-Yuan Hsiao, Puyuan Peng, Shih-Heng Wang, Chun-Yi Kuan, Ke-Han Lu, Kai-Wei Chang, Fabian Ritter-Gutierrez, Kuan-Po Huang, Siddhant Arora, You-Kuan Lin, Ming To Chuang, Eunjung Yeo, Kalvin Chang, Chung-Ming Chien, Kwanghee Choi, Jun-You Wang, Cheng-Hsiu Hsieh, Yi-Cheng Lin, Chee-En Yu, I-Hsiang Chiu, Heitor R. Guimarães, Jionghao Han, Tzu-Quan Lin, Tzu-Yuan Lin, Homu Chang, Ting-Wu Chang, Chun Wei Chen, Shou-Jen Chen, Yu-Hua Chen, Hsi-Chun Cheng, Kunal Dhawan, Jia-Lin Fang, Shi-Xin Fang, Kuan-Yu Fang Chiang, Chi An Fu, Hsien-Fu Hsiao, Ching Yu Hsu, Shao-Syuan Huang, Lee Chen Wei, Hsi-Che Lin, Hsuan-Hao Lin, Hsuan-Ting Lin, Jian-Ren Lin, Ting-Chun Liu, Li-Chun Lu, Tsung-Min Pai, Ankita Pasad, Shih-Yun Shan Kuan, Suwon Shon, Yuxun Tang, Yun-Shao Tsai, Jui-Chiang Wei, Tzu-Chieh Wei, Chengxi Wu, Dien-Ruei Wu, Chao-Han Huck Yang, Chieh-Chi Yang, Jia Qi Yip, Shao-Xiang Yuan, Vahid Noroozi, Zhehuai Chen, Haibin Wu, Karen Livescu, David Harwath, Shinji Watanabe, and Hung yi Lee. 2025a. Dynamic-superb phase-2: A collaboratively expanding benchmark for measuring the capabilities of spoken language models with 180 tasks.

Chien-Yu Huang, Ke-Han Lu, Shih-Heng Wang, Min-Han Shih, Chi-Yuan Hsiao, Chun-Yi Kuan, Haibin Wu, Siddhant Arora, Kai-Wei Chang, Jiatong Shi, Yifan Peng, Roshan Sharma, Shinji Watanabe, Bhiksha Ramakrishnan, Shady Shehata, and Hung yi Lee. 2024. Dynamic-superb: Towards a dynamic, collaborative, and comprehensive instruction-tuning benchmark for speech.

Chien-Yu Huang, Min-Han Shih, Ke-Han Lu, Chi-Yuan Hsiao, and Hung-Yi Lee. 2025b. Speechcaps: Advancing instruction-based universal speech models with multi-talker speaking style captioning. In *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5.

Yuxin Jiang, Yufei Wang, Xingshan Zeng, Wanjun Zhong, Liangyou Li, Fei Mi, Lifeng Shang, Xin Jiang, Qun Liu, and Wei Wang. 2023. Followbench: A multi-level fine-grained constraints following benchmark for large language models.

Chris Dongjoo Kim, Byeongchang Kim, Hyunmin Lee, and Gunhee Kim. 2019. AudioCaps: Generating captions for audios in the wild. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 119–132, Minneapolis, Minnesota. Association for Computational Linguistics.

You Jin Kim, Hee-Soo Heo, Soyeon Choe, Soo-Whan Chung, Yoohwan Kwon, Bong-Jin Lee, Youngki Kwon, and Joon Son Chung. 2021. Look who's talking: Active speaker detection in the wild. In *Interspeech 2021*, pages 3675–3679.

Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. 2023. Seed-bench: Benchmarking multimodal llms with generative comprehension.

Rensis Likert. 1932. A technique for the measurement of attitudes. *Archives of psychology*.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, Kai Chen, and Dahua Lin. 2024a. Mmbench: Is your multi-modal model an all-around player?

Yuanxin Liu, Shicheng Li, Yi Liu, Yuxiang Wang, Shuhuai Ren, Lei Li, Sishuo Chen, Xu Sun, and Lu Hou. 2024b. Tempcompass: Do video llms really understand videos?

Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. 2022. ChartQA: A benchmark for question answering about charts with visual and logical reasoning. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2263–2279, Dublin, Ireland. Association for Computational Linguistics.

Minesh Mathew, Dimosthenis Karatzas, and C. V. Jawahar. 2021. Docvqa: A dataset for vqa on document images.

Ivan Moshkov, Darragh Hanley, Ivan Sorokin, Shubham Toshniwal, Christof Henkel, Benedikt Schifferer, Wei Du, and Igor Gitman. 2025. Aimo-2 winning solution: Building state-of-the-art mathematical reasoning models with openmathreasoning dataset. *arXiv preprint arXiv:2504.16891*.

Arsha Nagrani, Joon Son Chung, and Andrew Zisserman. 2017. Voxceleb: A large-scale speaker identification dataset. In *Interspeech 2017*, pages 2616–2620.

Le Thien Phuc Nguyen, Zhuoran Yu, Samuel Low Yu Hang, Subin An, Jeongik Lee, Yohan Ban, SeungEun Chung, Thanh-Huy Nguyen, JuWan Maeng, Soochahn Lee, and Yong Jae Lee. 2025. See, hear, and understand: Benchmarking audiovisual human speech understanding in multimodal large language models.

OpenAI-Team. 2025. Introducing gpt-5.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. 2025. Qwen2.5 technical report.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R. Bowman. 2023. Gpqa: A graduate-level google-proof qa benchmark.

Maarten Sap, Hannah Rashkin, Derek Chen, Ronan LeBras, and Yejin Choi. 2019. Social iqa: Commonsense reasoning about social interactions. In *EMNLP*.

Yansong Tang, Dajun Ding, Yongming Rao, Yu Zheng, Danyang Zhang, Lili Zhao, Jiwen Lu, and Jie Zhou. 2019. Coin: A large-scale dataset for comprehensive instructional video analysis. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1207–1216.

Bin Wang, Xunlong Zou, Geyu Lin, Shuo Sun, Zhuohan Liu, Wenyu Zhang, Zhengyuan Liu, AiTi Aw, and Nancy F Chen. 2025a. Audiobench: A universal benchmark for audio large language models. *NAACL*.

Ke Wang, Junting Pan, Weikang Shi, Zimu Lu, Houxing Ren, Aojun Zhou, Mingjie Zhan, and Hongsheng Li. 2024a. Measuring multimodal mathematical reasoning with math-vision dataset. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Ke Wang, Junting Pan, Linda Wei, Aojun Zhou, Weikang Shi, Zimu Lu, Han Xiao, Yunqiao Yang, Houxing Ren, Mingjie Zhan, and Hongsheng Li. 2025b. Mathcoder-VL: Bridging vision and code for enhanced multimodal mathematical reasoning. In *The 63rd Annual Meeting of the Association for Computational Linguistics*.

Zirui Wang, Mengzhou Xia, Luxi He, Howard Chen, Yitao Liu, Richard Zhu, Kaiqu Liang, Xindi Wu, Haotian Liu, Sadhika Malladi, Alexis Chevalier, Sanjeev Arora, and Danqi Chen. 2024b. Charxiv: Charting gaps in realistic chart understanding in multimodal llms.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-thought prompting elicits reasoning in large language models.

Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. 2021. Next-qa:next phase of question-answering to explaining temporal actions.

Yuwei Yang, Zeyu Zhang, Yunzhong Hou, Zhuowan Li, Gaowen Liu, Ali Payani, Yuan-Sen Ting, and Liang Zheng. 2025. Effective training data synthesis for improving mllm chart understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

Xi Ye and Greg Durrett. 2022. The unreliability of explanations in few-shot prompting for textual reasoning.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. Hellaswag: Can a machine really finish your sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena.

Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. 2023. Instruction-following evaluation for large language models.

Nannan Zhu, Yonghao Dong, Teng Wang, Xueqian Li, Shengjun Deng, Yijia Wang, Zheng Hong, Tiantian Geng, Guo Niu, Hanyan Huang, Xiongfei Yao, and Shuaiwei Jiao. 2025. Cvbench: Evaluating cross-video synergies for complex multimodal understanding and reasoning.